A new feature extraction method for signal classification applied to cord dorsum potential detection

D Vidaurre , E E Rodríguez , C Bielza , P Larrañaga and P Rudomin

Abstract

In the spinal cord of the anesthetized cat, spontaneous cord dorsum potentials (CDPs) appear synchronously along the lumbo-sacral segments. These CDPs have different shapes and magnitudes. Previous work has indicated that some CDPs appear to be specially associated with the activation of spinal pathways that lead to primary afferent depolarization and presynaptic inhibition. Visual detection and classification of these CDPs provides relevant information on the functional organization of the neural networks involved in the control of sensory information and allows the characterization of the changes produced by acute nerve and spinal lesions. We now present a novel feature extraction approach for signal classification, applied to CDP detection. The method is based on an intuitive procedure. We first remove by convolution the noise from the CDPs recorded in each given spinal segment. Then, we assign a coefficient for each main local maximum of the signal using its amplitude and distance to the most important maximum of the signal. These coefficients will be the input for the subsequent classification algorithm. In particular, we employ gradient boosting classification trees. This combination of approaches allows a faster and more accurate discrimination of CDPs than is obtained by other methods.

1. Introduction

Classification of central nervous system signals recorded using different techniques, such as electrospinogram, electroencephalography or magnetoencephalography, is a task present in many biomedical scenarios, like, for example, brain–computer interface design [1–5].

Spontaneous spinal activity (SSA) in the cord dorsum was first recorded more than 60 years ago in [6] and [7] in the spinal cord of the cat. The SSA is characterized by a noise-like background activity recorded at the dorsal surface of the spinal cord. This can be observed as the occurrence of relatively large potentials in the absence of any stimulation (spontaneous cord dorsum potentials or CDPs). Studies using intact and freely moving cats show that the SSA recorded in such animals is similar to that observed in anesthetized animals except that frequency and amplitude are both lower [8].

Several investigators [9–11] have documented the nature of the spontaneous CDPs recorded in the anesthetized cat. These CDPs have different shapes and amplitudes. CDPs that are above 5–50 μV, start from a relatively flat baseline and last 40–70 ms, are usually purely negative CDPs (nCDPs) or negative–positive CDPs (npCDPs). The nCDPs and npCDPs appear to be generated by neurons located at the dorsal horn of the lumbar spinal cord, receiving mono and/or oligosynaptic

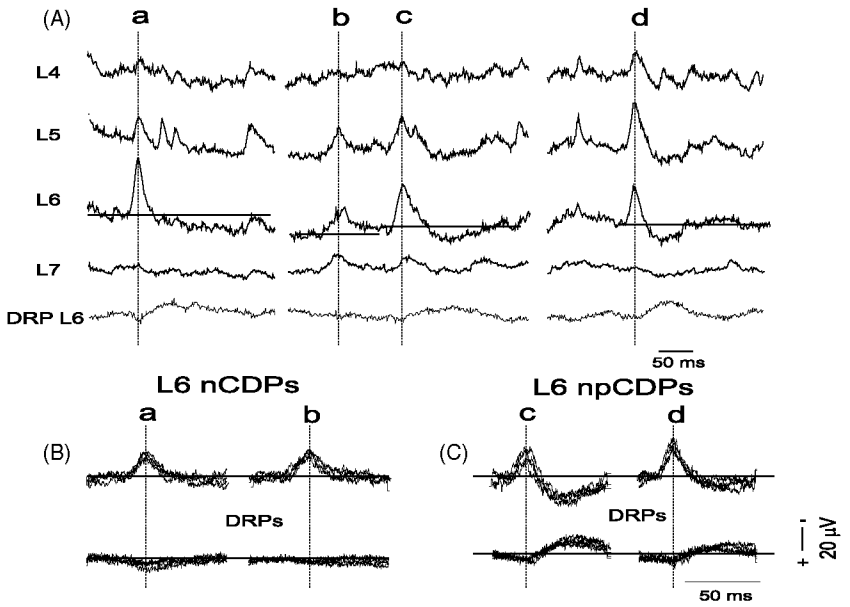

Figure 1. (A) Recordings that show synchronization and high correlation in different lumbar segments of CDPs (L4–L7) and DRPs of L6 (DRP L6) in the anesthetized cat. These occur throughout the recording at different time intervals. (B) nCDPs without DRP and npCDPs that appear to be associated with presynaptic inhibition due to the presence of DRP.

excitatory inputs from low-threshold cutaneous afferents [10]. See [12] for a review.

Unlike the nCDPs, the npCDPs have recently been found to be preferentially associated with spontaneous dorsal root potentials (DRPs), which are a sign of primary afferent depolarization and presynaptic inhibition [13]. Figure 1(A) illustrates samples of nCDPs and npCDPs recorded in one typical experiment that are similar to those reported in this paper. In this case, spontaneous CDPs were simultaneously recorded from four spinal segments on the left side (L4, L5, L6 and L7), together with the DRPs, recorded from the central end of a small L6 dorsal root filament on the left side. The CDPs recorded from the L6 segment were usually larger than those recorded from the other segments. Vertical lines labeled a, c and d show three L6 npCDPs associated with DRPs, while line b shows nCDPs occurring without DRPs. Panels B and C show several superposed nCDPs, npCDPs and corresponding DRPs. Note that, in C, the npCDPs with a larger positive component appeared in association with the largest DRPs.

In previous studies, the program used to separate the spontaneous CDPs according to their shape and amplitude was based on the visual selection of a few nCDPs and npCDPs whose means were later used as templates to retrieve the nCDPs and npCDPs for the whole recording period [13]. Usually three experts visually inspected the output CDPs to remove signals without the predetermined characteristics (see [13] for more details). The visual selection of nCDPs and npCDPs took hours or even days. Therefore, we aimed at designing a faster and automatic procedure to classify spontaneous CDPs.

Typically, some feature extraction approach applied to raw signals precedes the classification procedure. Some well-known examples are based on amplitude values [14], band powers [15], power spectral density values [16], autoregressive coefficients (AR) [17], principal component analysis (PCA) [18] and independent component analysis (ICA) [19].

In this paper, we propose to analyze the main peaks of the CDPs and summarize the entire signal in a few coefficients derived from the amplitude and separation of the peaks. The objective is to mimic the intuitive classification rule used by the experts to distinguish spontaneous CDPs generated by neurons located in the dorsal horn of the lumbar spinal cord.

The rest of the paper is organized as follows. Section 2 presents the underlying methodology on which the proposed approach is based. In particular, we give a brief description of the discrete wavelet transform (DWT), boosting classification trees and state-of-the-art feature extraction. Section 3 introduces the core of the method. Sections 4 and 5 describe data and results with, respectively, synthetic and real data sets. Finally, section 6 sums up the paper and outlines future work.

2. Underlying methodology

In this section, we briefly describe the DWT [20], which we use to approximate the signal while removing noise. Also, we outline the principles of boosting classification trees [21]. Although we have tested other supervised classification paradigms, such as bagged trees, support vector machines or quadratic discriminant analysis [21], they have been found to behave worse in this setting.

2.1. Discrete wavelet transform

Unlike Fourier analysis, which establishes a frequency representation of an analog signal, wavelet theory uses a time–frequency representation. The DWT maps an input signal of T values onto components of different frequencies. For

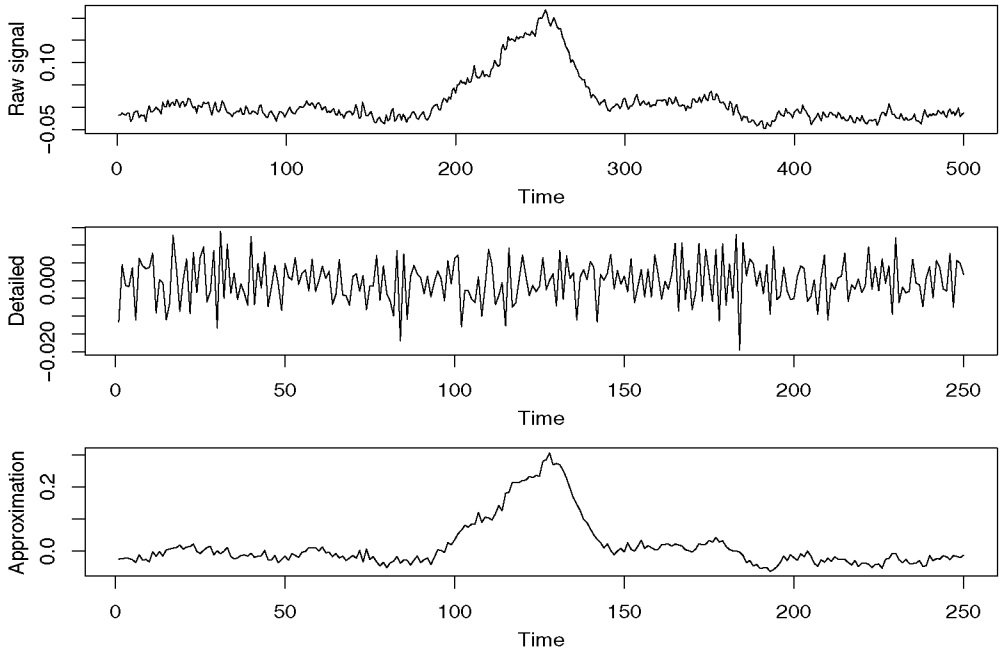

Figure 2. Wavelet signal decomposition of some signal (top) into a detailed component (middle) and an approximation component (bottom).

simplicity, T is usually considered to be a power of 2, but this is not strictly necessary.

Let $\boldsymbol{x} = (x_1, \ldots, x_T)'$ be a discretized signal. The DWT of $\boldsymbol{x}$ is computed by passing it through a sequence of filters. The signal is decomposed using a high-pass filter and a low-pass filter. In this paper, we use a least-asymmetric mother wavelet filter of length 8. After downsampling the redundant information, this process halves the time resolution, splitting the signal into two vectors of $T/2$ values: the detailed $\boldsymbol{x}^0$, generated by the high-pass filter, and the approximation $\boldsymbol{x}^1$, generated by the low-pass filter. Figure 2 illustrates a $T = 250$ raw signal (top) decomposed into a detailed component (middle) and an approximation component (bottom).

We can further decompose $\boldsymbol{x}^1$ down to level p, obtaining an approximation vector $\boldsymbol{x}^p$ with $T/2^p$ coefficients. Thus, $\boldsymbol{x}^p$ contains a noise-free, compact description of the original signal $\boldsymbol{x}$, whose detail level depends on p. As explained below, this is the first step of our approach. We use the `wavelets` R package[4] for this purpose.

2.2. *Boosting classification trees*

Boosting is an extremely successful idea within machine learning theory. In this paper, we apply boosting classification trees to the preprocessed signals, in the final classification step. Boosting classification trees are based on the combination of many simple base learners to produce a powerful final classifier. The base learners are simple classification trees. Prediction is performed by a weighted majority vote:

$$\mathscr{T} = \text{sign}\left(\sum_{\text{k=1}}^{\text{K}} \text{w}_\text{k}\mathscr{T}^{(\text{k})}\right), \tag{1}$$

where the actual classification $\mathscr{T}$ is the weighted sum (given weights w_k) of the outputs $\mathscr{T}^{(k)} \in \{-1, 1\}$ of K (sequentially built) simple classification trees. This scheme can be generalized by using base learners that output real-valued confidence predictions (probabilities) mapped to the interval $[-1, 1]$ [22]. To keep the notation uncluttered, we drop the explicit algorithm input from the expression.

Hence, at each step k, the algorithm has to induce both a classification tree and a weight w_k. The classification tree is trained over a weighted version of the data set, using the weights $\nu_1, \ldots, \nu_N$, where N is the number of data instances. The weights $\nu_1, \ldots, \nu_N$ are computed so that data instances that were misclassified in the previous age are given more importance. The weights w_k are computed as a function of the error at this step, typically as $w_k = \log((1 - \epsilon_k)/\epsilon_k))$, where ϵ_k is the error made by the tree k. In this way, more influence is attached to the more accurate trees.

Following this scheme, the boosting procedure can be very efficiently performed using a two-terminal node classification tree (also called decision stump) as the basic classifier; see [21] for details. The method works insofar as single classifiers are (just slightly) better than random guessing, which is the case for two-terminal node classification trees. As it uses only two-terminal node trees, however, the resulting model does not consider interactions between the variables. As a general rule, if $(J + 1)$-terminal node trees $(J \geqslant 1)$ are used, the final model considers interactions up to order J. Here, we consider two-level interactions.

In this paper, we use a faster, more sophisticated implementation of this basic scheme, called gradient boosting [23]. Gradient boosting makes use of numerical optimization techniques to approximate the solution of the boosting classification trees problem (in this case, with $J = 2$). In particular, the minimization of the binomial likelihood loss function derived from equation (1) can be formulated as a

[4] http://cran.r-project.org/web/packages/wavelets/index.html.

convex optimization problem and solved by gradient descent. We use the `gbm` R package[5].

Since the training data set is formed by a collection of signals whose class (CDP or not CDP) is determined by visual identification, it is natural to consider a richer representation of the class that reflects label uncertainty. This is motivated by the error proneness and subjectivity of the labeling process. For example, instead of a binary value, the algorithm input can be a percentage reflecting the confidence of a signal being a CDP, mapped to be in the interval $[-1, 1]$. If several experts are available, this percentage can be obtained from the ratio of experts that labeled the signal as a CDP. We can use gradient boosting trees to handle uncertainty in the labeling process by applying the boosting regression trees approach. This approach minimizes a squared error loss function instead of the binomial likelihood. Since the actual prediction is a real value in $[-1, 1]$, we can use the sign for the actual classification. The `gbm` R package includes the boosting regression trees methodology. A natural alternative, not implemented in `gbm`, is to adapt the binomial likelihood loss function to make use of CDP probabilities instead of binary responses.

2.3. State-of-the-art feature extraction methodology

In this section, we briefly describe three popular feature extraction techniques that will be used in sections 4 and 5 for comparison purposes: AR [17], PCA [24] and ICA [25].

The AR method is based on linear regression, where the responses are the signal values and the regression covariates are the γ previous signal values. Assuming centered data, a γ-order AR model is a type of random process that imposes a linear relation

$$x_t = \sum_{l=1}^{\gamma} \omega_l x_{t-l} + \varepsilon, \quad t \in \{\gamma + 1, \ldots, T\},$$

where ε is the Gaussian white noise and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_\gamma)'$ are the AR coefficients, which will play the role of the predictors in the subsequent classification algorithm. The tuning parameter is thus γ.

PCA obtains a linear decomposition of the data intended to capture maximal variance. Let $\boldsymbol{X}$ denote the $N \times T$ matrix with one row per signal. PCA is based on the eigen-decomposition of the sample covariance matrix $\boldsymbol{X}'\boldsymbol{X}/N$, defined as

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}',$$

where $\boldsymbol{V}$ is a $T \times T$ orthogonal matrix spanning the row space of $\boldsymbol{X}$ (an orthogonal basis) and $\boldsymbol{D}$ is a $T \times T$ diagonal matrix, with diagonal entries $d_1 \geqslant d_2 \geqslant \ldots \geqslant d_p$. Such values are the singular values of $\boldsymbol{X}$. The classifier inputs are the first Q columns of $\boldsymbol{V}$ that correspond to the highest eigenvalues, i.e. the columns which have the highest variance among all the linear combinations of the data set columns. Those columns are called the principal components. The tuning parameter can be either the percentage of variance that we intend to capture or the value of Q. We use the `prcomp` built-in R command.

ICA aims at separating the different sources from which some multivariate data are generated, identifying a matrix of independent latent components that we can use as the classification algorithm input. The difference from classic factor analysis [26] is that ICA is built under the assumption of mutual statistical independence and non-Gaussianity of the sources, whereas factor analysis assumes non-correlated, Gaussian distributed data. In this paper, we run ICA on the projection of $\boldsymbol{X}$ onto its Q principal component directions, i.e. on the first Q columns of $\boldsymbol{V}$, previously computed by PCA. Let $\boldsymbol{V}_Q$ be the $T \times Q$ matrix with the first Q columns of $\boldsymbol{V}$. The ICA model is defined as

$$\boldsymbol{X}\boldsymbol{V}_Q = \boldsymbol{S}\boldsymbol{A}',$$

where $\boldsymbol{A}$ is a $Q \times Q$ orthogonal matrix of loadings and $\boldsymbol{S}$ is an $N \times Q$ matrix that encodes the latent variables or factors, which represent common sources of variation for $\boldsymbol{X}\boldsymbol{V}_Q$. The columns of $\boldsymbol{S}$ represent non-Gaussian, independent variables. It is assumed that $\boldsymbol{V}_Q'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}_Q = N\boldsymbol{I}$ and $\boldsymbol{S}'\boldsymbol{S} = N\boldsymbol{I}$. The objective is to find $\boldsymbol{A}$ such that $\boldsymbol{S}$ holds the mentioned conditions. $\boldsymbol{A}$ is typically estimated by information theory techniques, such as the minimization of the mutual information between the components of $\boldsymbol{X}\boldsymbol{V}_Q\boldsymbol{A}$. When the estimates are constrained to be uncorrelated, this amounts to maximizing the departure from Gaussianity of the estimates. Then, the tuning parameter is Q and the columns in $\boldsymbol{S}$ are the extracted features. We use the `fastICA` R package.

3. Feature extraction based on peak analysis

Our aim is to use some feature extraction method to map each T-value signal onto a meaningful vector of M components, where M is some small value. A gradient boosting algorithm is then run to train an accurate classifier on these M-value vectors. M is, then, the number of features to extract. The general idea is to represent each main peak (either maximum or minimum) by some value that quantifies its magnitude and distance to the main maximum. In the following, we detail each step of the devised feature extraction procedure[6]. This is enacted separately for each signal.

The first step is to approximate each signal by the DWT (see section 2.1) in order to retain the main information and then identify the peaks. The DWT transforms each original T-value signal into an approximation $\boldsymbol{s} = (s_1, \ldots, s_{T/2^p})'$. We denote the signal value of peak i as $\alpha_i \in \{s_1, \ldots, s_{T/2^p}\}$, placed at the time point $t_i \in \{1, \ldots, T/2^p\}$. Obviously, if α_i is a minimum, then α_{i-1} and α_{i+1} correspond to maxima and vice versa (unless α_i is the leftmost or rightmost peak).

Let $\Delta_{\max}$ be the height difference between the highest maximum and the lowest minimum, $\Delta_{\max} = \max\{|\alpha_i - \alpha_{i'}|\}_{i \neq i'}$, where $\{|\alpha_i - \alpha_{i'}|\}_{i \neq i'}$ is the set of height differences between all peaks of the approximated signal. We assume that the signal is not completely flat and has at least one peak, because, otherwise, the signal would not have been considered. Let $t_{\max}$ be the time point of the highest maximum.

We discard all peaks α_i that do not satisfy

$$|\alpha_i - \alpha_{i-1}| \geqslant \delta\Delta_{\max} \quad \text{and} \quad |\alpha_i - \alpha_{i+1}| \geqslant \delta\Delta_{\max}, \tag{2}$$

[5] http://cran.r-project.org/web/packages/gbm/index.html.

[6] R code is available on request.

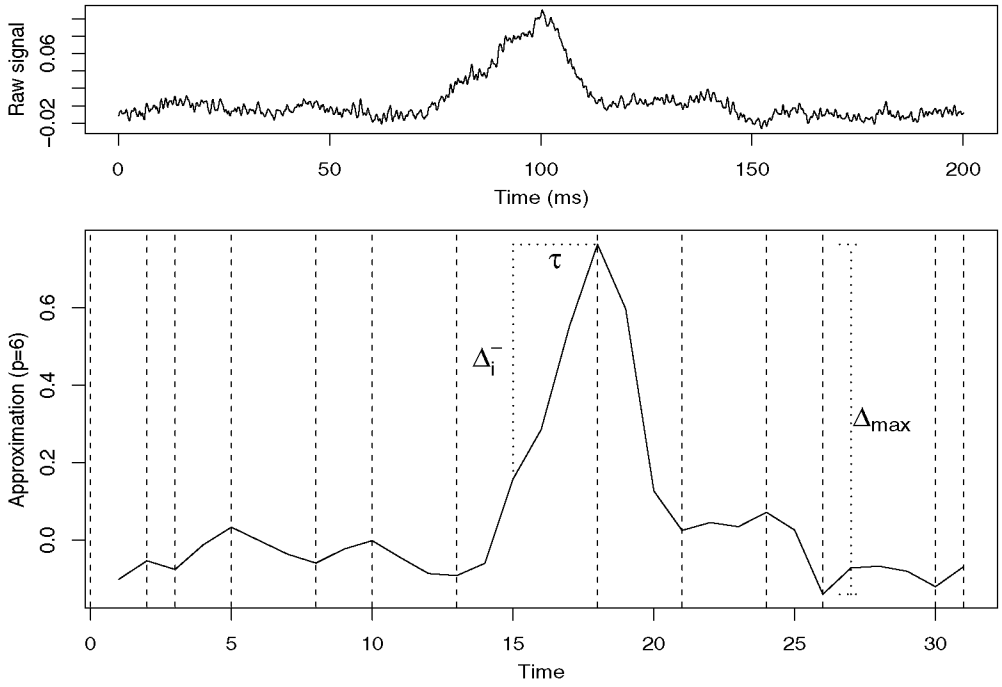

Figure 3. Raw 200 ms signal (top) and $p = 6$ approximation component (bottom), where main peaks are indicated by vertical lines. Only the mean peak is marked in the top signal. The time gap between adjacent points in the bottom graph is 3.2 ms.

where $\delta \in (0, 1)$ is some parameter. This way, we discard those peaks that do not have a minimum slope. Figure 3 illustrates a raw signal (top) and the $p = 6$ approximation component, where the selected peaks (peaks that satisfy condition (2)) are indicated by dashed lines (bottom).

Since we want to characterize the signal according to the nature of its maxima, the next step is to assign a measure β_i to each *maximum*. Assuming that α_i is a maximum, we define β_i as

$$\beta_i = \Delta_i^- \, \Delta_i^+ \, \phi_\sigma(t_i - t_{\max}), \tag{3}$$

where Δ_i^- and Δ_i^+ are defined, given some parameter τ, as

$$\Delta_i^- = \begin{cases} \max\{|\alpha_i - s_j|\}_{t_i - j \leqslant \tau} & \text{if } t_i - t_{i-1} \geqslant \tau \\ |\alpha_i - \alpha_{i-1}| & \text{otherwise} \end{cases}$$

and

$$\Delta_i^+ = \begin{cases} \max\{|\alpha_i - s_j|\}_{0 < j - t_i \leqslant \tau} & \text{if } t_{i+1} - t_i \geqslant \tau \\ |\alpha_i - \alpha_{i+1}| & \text{otherwise,} \end{cases}$$

and $\phi_\sigma(\cdot)$ is some kernel function. We use the well-known tri-cube kernel

$$\phi_\sigma(d) = \begin{cases} (1 - |d|^3)^3 & \text{if } |\mathrm{d}| \leqslant \sigma \\ 0 & \text{otherwise,} \end{cases}$$

where $\sigma > 0$ is the kernel parameter.

Intuitively, Δ_i^- (Δ_i^+) measures the signal difference between this maximum and the minimum just on the left (on the right). If this minimum is further than τ, then Δ_i^- (Δ_i^+) is the signal decrement within this τ-radius neighborhood. The value Δ_i^- for the principal maximum is indicated for $\tau = 3$ in figure 3.

We now keep the $M + 1$ highest β_i values. Recall that we have defined β_i only when α_i is a maximum. We form the $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_{M+1}^*)'$ vector by sorting the defined elements β_i in decreasing order:

$$\beta_1^* > \beta_2^* > \ldots > \beta_{M+1}^*,$$

where the last components can be zero.

Algorithm 1. Feature extraction based on peak analysis.

Obtain an approximation $\boldsymbol{s}$ with the DWT.
Identify the peaks and store their signal values in $\boldsymbol{\alpha}$.
Compute $\Delta_{\max} = \max(\{\alpha_i - \alpha_{i'}\}_{i \neq i'})$.
Discard those peaks that do not satisfy equation (2).
Compute $\boldsymbol{\beta}$ by equation (3).
Sort $\boldsymbol{\beta}$ and keep the $M + 1$ first components to obtain $\boldsymbol{\beta}^*$.
Output $(\beta_2^*/\beta_1^*, \ldots, \beta_{M+1}^*/\beta_1^*)'$.

Finally, we rescale $\boldsymbol{\beta}^*$ by applying the perspective function, which divides each element of $\boldsymbol{\beta}^*$ by β_1^*, and we then remove the first component from the resulting vector (which is equal to 1). The resulting vector of inputs for the supervised classifier is then $(\beta_2^*/\beta_1^*, \ldots, \beta_{M+1}^*/\beta_1^*)'$. The entire procedure is repeated for each signal in the data set. Algorithm 1 outlines the proposed feature extraction procedure.

Table 1 shows the value of β_i, obtained with equation (3), for the maxima of the signal depicted in figure 3. Note that the minima are not assigned a β_i value. Note also that some peaks (peaks that correspond to time points 21, 22 and 28 in the approximation scale) have been discarded because they do not satisfy equation (2). For $M = 3$, for example, we would have $\boldsymbol{\beta}^* = (0.4468, 0.0090, 0.0042, 0.0038)'$.

Note that, if the signal has two relevant peaks, the method will consider the highest peak as the maximum peak, treating the other one as any other peak. If two or more peaks are very

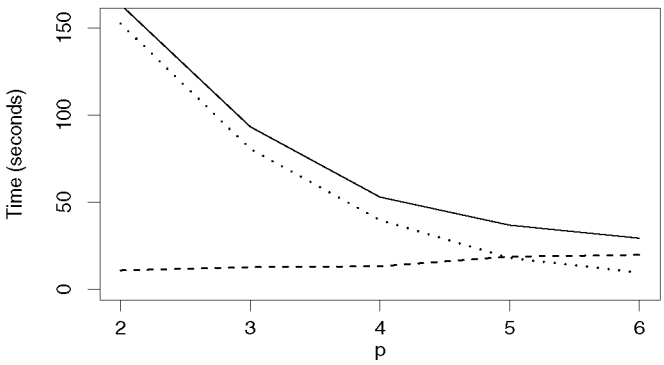

Figure 4. Time in seconds taken by the entire feature extraction method (solid line), showing the computation time of the DWT phase (dashed line) and the remaining steps (dotted line).

Table 1. Time point (t_i), signal value (α_i), type of peak and values of Δ_i^-, Δ_i^+ and β_i of the selected peaks for the signal depicted in figure 3. The main peak is highlighted.

i	t_i	α_i	Type	Δ_i^-	Δ_i^+	β_i
1	2	−0.05	Maximum	0.0476	0.0223	0.0001
2	3	−0.08	Minimum	–	–	–
3	5	0.03	Maximum	0.1090	0.0925	0.0038
4	8	−0.06	Minimum	–	–	–
5	10	0.00	Maximum	0.0578	0.0902	0.0042
6	13	−0.09	Minimum	–	–	–
7	**18**	**0.76**	**Maximum**	**0.6054**	**0.7379**	**0.4468**
8	21	0.03	Minimum	–	–	–
11	24	0.07	Maximum	0.0467	0.2114	0.0090
12	26	−0.14	Minimum	–	–	–
14	30	−0.12	Minimum	–	–	–

close to each other, then, by definition, they will not be CDPs, and, then, the signal will be labeled as non-CDP. Otherwise, if the signal is considered as a CDP, then the 'main' peak is expected to be relatively far from the others. Consequently, thanks to the kernel function $\phi_\sigma(\cdot)$, the β_i value of the second peak will be low and will probably not be included in the vector of extracted features.

The proposed method is fast thanks to the DWT step, with parameter p, which shortens the signals to handier sizes. The DWT is a rapid procedure, too. For example, figure 4 shows, for different values of p, the time in seconds taken by the entire feature extraction method for a real data set (described below) with $N = 1577$ signals and $T = 2001$ (200 ms), divided into the DWT phase and the remaining steps. We have used an Intel Core 2 Duo processor (2.26 GHz). The standard deviation of time across different executions is negligible. Note that p is the only parameter related to computational efficiency. Classification by gradient boosting is also very fast for moderate dimensions.

In summary, the proposed feature extraction approach involves the configuration of several input parameters: p, M, δ, τ and σ. Although a cross-validation procedure can be enacted to select all these parameters, most of them are not crucial and can be set to any reasonable value. Note that, given the computational efficiency of the method, cross-validation is affordable. The relevant parameters are, in fact, M and p. The choice of p should be governed by the signal length and the signal-to-noise ratio. For longer and noisier signals, the value of p is incremented in order to, respectively, reduce the computational burden and avoid spurious minima/maxima. M can be selected by cross-validation. For the other parameters, we have observed that the following values yield good results: $\delta = 0.01$, $\tau = 3$ and $\sigma = 18$. The classification output does not vary much for sensible variations of these parameters.

4. Experiments with synthetic data

4.1. Data generation

Synthetic data were obtained by adding several types of noise (with different frequencies and amplitudes) to a pure CDP without noise (like, for example, those in figure 1(C)) Each signal is thus labeled with its confidence level being a CDP (either nCDP or npCDP), which depends on the type and amount of added noise. The confidence level is a quantitative measure of how well each signal fits the definition of CDP according to an external observer. Signals with 1.0 CDP confidence have only Gaussian white noise with low variance. We have added a low-amplitude noise sinusoidal signal to signals with 0.8 CDP confidence. In addition to the low-amplitude noise signal, we have added a medium-amplitude noise sinusoidal signal to signals with 0.6 CDP confidence. Also, we have added a high-amplitude noise signal to signals with 0.4 CDP confidence. Signals with 0.0 CDP confidence have only noisy sinusoidal signals and no CDP signal at all. Finally, each signal was translated at random over the time scale. Figure 5 shows some examples of generated signals. The generated data set comprises $N = 500$ signals with $T = 2000$ (200 ms) time points. Of these, 80 are 1.0 confidence CDPs, 80 are 0.8 confidence CDPs, 80 are 0.6 confidence CDPs, 80 are 0.4 confidence CDPs and 180 are non-CDPs (with a 0.0 confidence level). Note that the objective of these experiments is to measure how the models are capable of learning from data and not to give a rigorous CDP characterization. Hence, the models obtained from this data set should not be used for classification of future real signals.

4.2. Results

In this section, we compare the feature extraction approach based on peak analysis (PA) with AR, PCA and ICA on a synthetic data set. We have run the gradient boosting classification trees algorithm on the features extracted by AR, PCA, ICA and PA. We also test the amplitude thresholding (AT) method, which is widely used, for example, for neural spike detection [27]. In an unsupervised manner, AT would select the signals whose main peak amplitude exceeds a certain threshold, which is typically set as a multiple of the estimated noise standard deviation. In this case, since we have a labeled training data set, we obtain the median of the main peak amplitudes (normalized by the estimated noise standard deviation) separately for CDPs and non-CPDs. We denote them, respectively, as m_{CDP} and m_{nonCDP}. In a supervised way, an incoming signal will be classified as CDP when its normalized main peak amplitude is closer to m_{CDP}

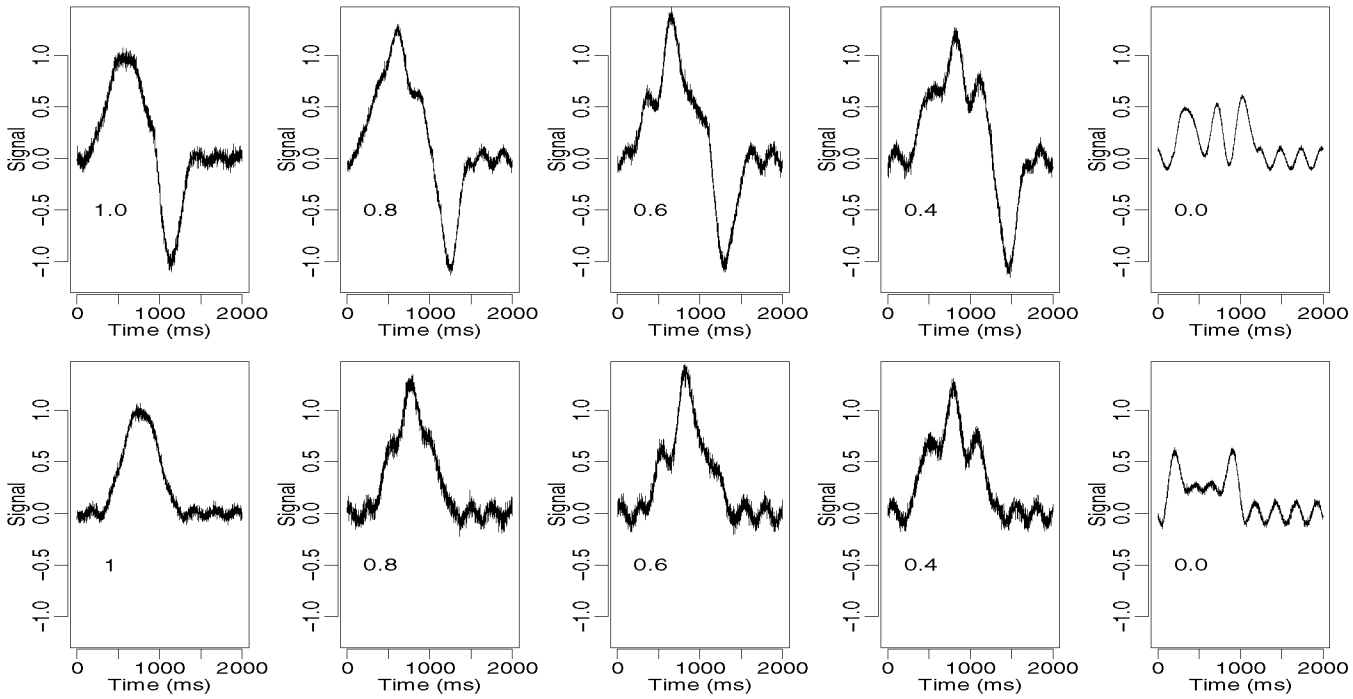

Figure 5. Some artificially generated signals. Each chart indicates the confidence level of the signal being considered a CDP (either nCDP (top) or npCDP (bottom)).

than to m_{nonCDP}. Then, the threshold is automatically set to $(m_{\text{CDP}} + m_{\text{nonCDP}})/2$. Unlike PA, AT does not consider the relation between peaks, exclusively focusing on the main peak. Before applying AR, PCA, ICA and AT, a low-pass filter with detail level $p = 2$ is applied to each signal to remove noise, so that signals have 500 time points.

For our approach, we have set $p = 6$, so that the signals are reduced from $T = 2000$ to 32 time points. Also, we have used $\delta = 0.01$, $\tau = 3$ and $\sigma = 18$. We have performed 20-fold cross-validation to evaluate the methods, leaving 25 signals out for testing at each iteration and using the remaining 475 for training and selection of the parameters M (PA), γ (AR), Q (PCA) and L (ICA). Note that AT does not require parameter selection, so, at each cross-validation iteration, we use the 475 samples for training.

Table 2 shows mean and standard deviations of sensitivity and specificity for each approach. Sensitivity is the number of identified CDPs divided by the total number of true CDP. Specificity is the number of identified non-CDP signals divided by the total number of signals that are not CDPs. For evaluation purposes, a signal is considered to be a CDP if the confidence is higher than 0.5.

On the one hand, PA, AR and PCA yield a high sensitivity. Although sensitivity is best for PCA, the differences to PA and AR are not statistically significant according to a t-test with a significance level of 0.01. The t-test, following Student's t-distribution, has been arranged to be one-sided. As expected, since it only considers one peak and, hence, disregards the relation between peaks, AT exhibits a poor sensitivity. On the other hand, PA clearly yields the highest specificity, with a statistically significant difference from the other methods. Note that PCA, along with AR, has the worst specificity, overshadowing the good sensitivity results. ICA, instead, gives a fine balance between sensitivity and specificity, but this is still worse than PA. PA, then, clearly produces the best sensitivity–specificity compromise.

Figure 6 illustrates the accuracy for each method and each CDP confidence level. For each confidence level in the x-axis, we show the results for the signals with such confidence level. For evaluation purposes, we consider a signal to be a CDP if its confidence is greater than 0.5. The figure is divided by a dotted vertical line, so that the left part corresponds to signals with confidence level lower than 0.5 (considered non-CDP) and the right part corresponds to signals with confidence level greater than 0.5 (considered CDP). Whereas the accuracy reported on the left part of the chart accounts for specificity errors (non-CDP signals misclassified as CDPs), the accuracy reported on the right accounts for sensitivity errors (CDP signals misclassified as non-CDPs). Note that PA is the only method that is able to correctly classify CDP with an uncertain label (0.4 and 0.6). This can be interpreted as a sign of robustness.

5. Experiments with real data

5.1. Data acquisition

Data were obtained from four control recordings performed in adult cats. Guidelines contained in Principles of Laboratory Animal Care[7] were followed in all cases and the experiments were also approved by the Institutional Bioethical Committee[8]. The animals were initially anesthetized with pentobarbitone sodium (40 mg kg^{-1} i.p.) and additional doses were given

[7] NIH publications 85-23, revised in 1985.
[8] Protocol number: 0126-03.

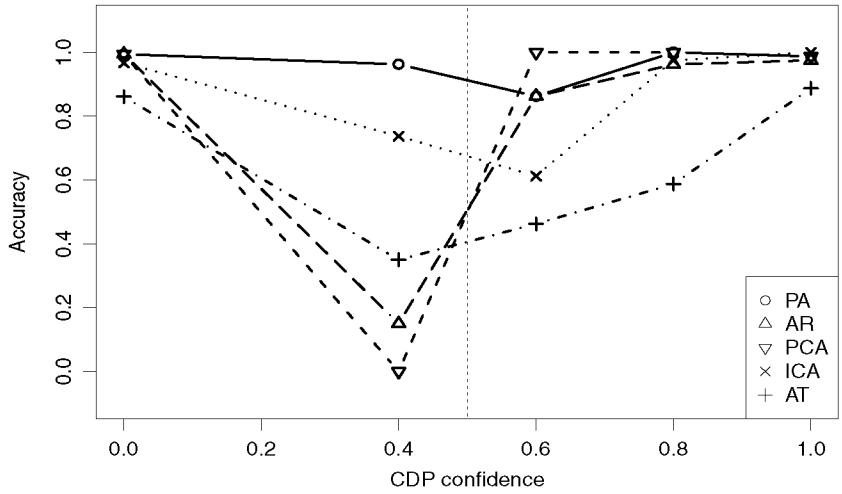

Figure 6. Accuracy for each method and each CDP confidence level.

Table 2. Sensitivity and specificity results for the boosting classification trees algorithm after feature extraction performed by PA, AR, PCA and ICA, and for AT, for the synthetic data set. The best results are highlighted. The symbol $*$ is added when the difference between the best and the second-best method is statistically significant.

	PA	AR	PCA	ICA	AT
Sensitivity	0.95(±0.07)	0.94(±0.06)	**0.99**(±0.09)	0.86(±0.10)	0.65(0.22±)
Specificity	**0.98** $*$ (±0.03)	0.74(±0.09)	0.68(±0.12)	0.89(±0.08)	0.86(0.12±)

intravenously to maintain an adequate level of anesthesia, tested by assessing that withdrawal reflexes were absent, the pupils were constricted and the arterial blood pressure was between 100 and 120 mm Hg^{-1}. The carotid artery, radial vein, trachea and urinary bladder were cannulated. A solution of 100 mM of sodium bicarbonate with glucose 5% was given i.v. (0.03 ml min^{-1}) to prevent acidosis [28]. When necessary, dextran 10% or ethylephrine (Effortil, Boering-Ingelheim) was administered to keep blood pressure above 100 mm Hg^{-1}.

The lumbosacral and low thoracic spinal segments were exposed. After the surgical procedures, the animals were transferred to a stereotaxic metal frame allowing immobilization of the spinal cord, paralyzed with pancuronium bromide (0.1 mg kg^{-1}) and artificially ventilated. The tidal volume was adjusted to maintain 4% of CO_2 concentration in the exhaled air. To prevent desiccation of the exposed tissues, pools were made with the skin flaps, filled with paraffin oil and maintained between 36° and 37° by means of radiant heat.

Four ball Ag–AgCI homemade electrodes were placed on the cord dorsum of the lumbosacral enlargement at different spinal segments to record the spontaneous CDPs against an indifferent electrode placed on the paravertebral muscles. The whole recording period of the SSA ranged from 10 to 30 min in order to have sufficient samples for further analysis. Spontaneous CDPs were recorded with separate preamplifiers (band-pass filters from 0.3 Hz to 10 kHz), displayed on an oscilloscope, digitized with a sampling rate of 10 kHz and stored for subsequent processing. After the experiment, the spontaneous CDPs recorded in L6 segments that exceeded a predetermined amplitude (5 μV) were sequentially displayed and aligned by centering the signal at the highest point of the L6 recordings. This way, we obtained $N = 1577$ signals, whose segment size is $T = 2001$ time points (200 ms), with 100 ms before the peak and 100 ms after the peak. Of these, only 379 are CDPs (where 210 are nCDPs and 169 are npCDPs) that could be associated with presynaptic mechanisms (npCDPs). No confidence levels are available for these data.

5.2. Results

In this section, we evaluate the described methods on real data. The tested techniques are again PA, AR, PCA, ICA and AT. We used a similar experimental scheme to that in section 4.2. In this case, however, to improve the algorithm's performance, after the feature extraction step, we have grown the data set by including (randomly selected) repeated copies of CDPs. This is done to balance the data set, because there are few CDP instances. Again, we use 20-fold cross-validation. Table 3 shows mean and standard deviations of sensitivity and specificity for each approach.

PA clearly produces the best sensitivity–specificity compromise, followed by ICA. Sensitivity is of special interest because the value of identifying the true CDPs is high. Note that sensitivity is the highest for PA, followed by PCA, from which the difference is statistically significant according to a t-test with a significance level of 0.01. PA and AR also have the best specificity. Note that, on the other hand, sensitivity is poor for AR and specificity is poor for PCA. The AT results are rather average.

Figure 7 shows some examples of signals that only PA is able to identify as CDPs (either an nCDP or an npCDP). They have been approximated with the DWT for clarity. Although

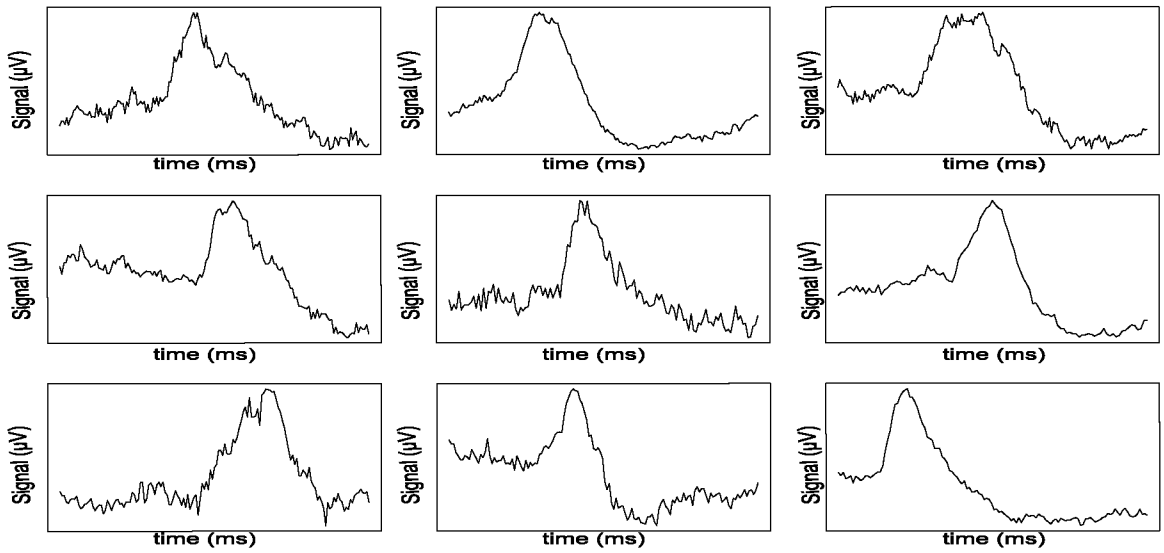

Figure 7. Some CPDs that were only identified by the PA approach. Most of them are npCDPs. Time scale is in ms and signal scale is in μV.

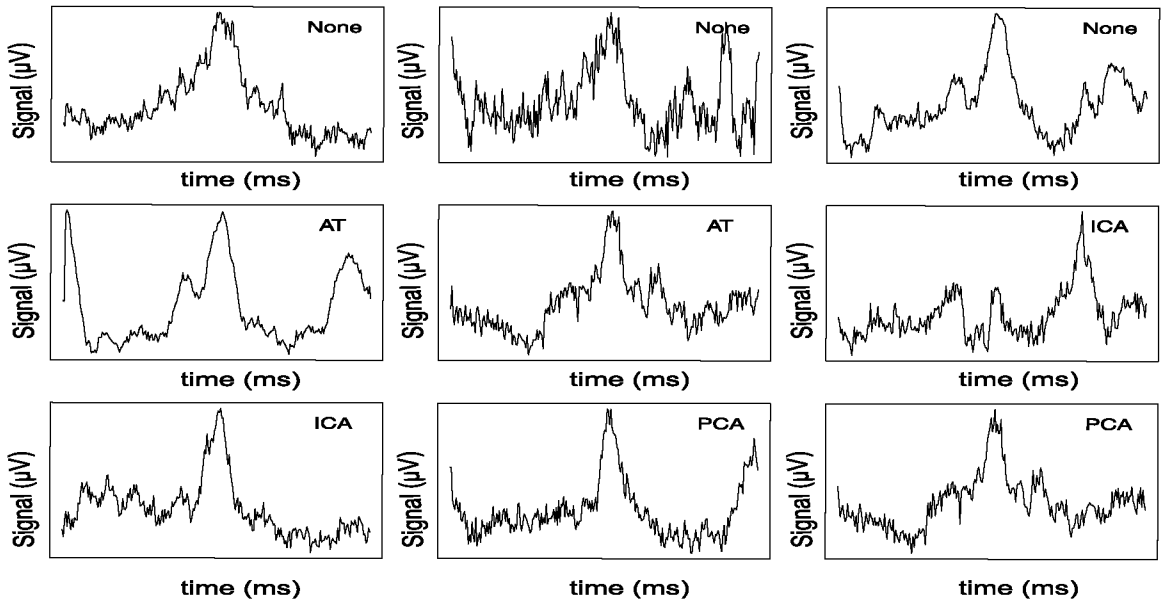

Figure 8. Some CDPs that could not be identified by the PA approach. The method that identified the signal as CDP (or none) is specified in each case. Time scale is in ms and signal scale is in μV.

Table 3. Sensitivity and specificity results for the boosting classification trees algorithm after feature extraction performed by PA, AR, PCA and ICA, and for AT, for the real data set. The best results are highlighted. The symbol $*$ is added when the difference between the best and the second-best method is statistically significant.

	PA	AR	PCA	ICA	AT
Sensitivity	**0.83** $*$ ($\pm$0.10)	0.54($\pm$0.14)	0.72($\pm$0.11)	0.67($\pm$0.09)	0.66($\pm$0.09)
Specificity	**0.67**($\pm$0.08)	**0.67**($\pm$0.04)	0.57($\pm$0.07)	0.64($\pm$0.06)	0.62($\pm$0.06)

these are not the best defined CDPs, they meet the requirements for categorization as CDPs. None of the other state-of-the-art feature extraction methods (or AT), however, lead to their identification.

Figure 8 shows some CDP signals obtained from the same set of recordings that PA could not identify. As observed, most of them are noisy or contain several significant peaks that are relatively close to the main peak. Although they were visually labeled as CDPs, their inclusion in this category is debatable because they do not have a well-defined baseline or they have multiple peaks around the main one. The signal in the middle row and left column, for example, has three

significant peaks. The proposed approach can thus be used to check for errors in the visual classification, and is, in any case, a handy aid for the slower manual procedure.

6. Discussion

In this paper, we have presented a novel feature extraction approach based on the signal peaks. We have verified the usefulness of the method for identifying spontaneous CDPs (nCDPs and npCDPs together in the same classification) that are generated by neurons located at the dorsal horn of the lumbar spinal cord receiving mono and/or oligosynaptic excitatory inputs from low-threshold cutaneous afferents. Experiments with synthetic data reinforce this claim. The algorithm is very fast and outperforms state-of-the-art methods for CDP recognition (nCDPs and npCDPs) in terms of accuracy. Hence, the introduced approach is a useful tool for preselecting well-defined CDPs and is an aid for a highly time-consuming manual procedure.

Note that some problem-related heuristics could be applied to further improve classification accuracy. For example, a significant maximum that is relatively close to the main maximum of the signal is discarded as a CDP with possible presynaptic inhibition association. Such heuristics have been excluded from the procedure for the sake of generality.

The method is not well suited for discriminating between nCDPs and npCDPs. Such automatic classification could be done by comparing the initial and final voltage values to ascertain the presence or absence of a positive component following the negative part. As mentioned above, nCDPs and npCDPs can also be distinguished by observing that, unlike the spontaneous nCDPs, the spontaneous npCDPs occur in association with DRPs, which are a sign of primary afferent depolarization and presynaptic inhibition. DRPs recordings are, however, not feasible in humans. Since presynaptic inhibition has been shown to be altered after spinal lesions (leading, for example, to spasticity, paresthesias and weakness [29, 30]), reliable discrimination of nCDPs and npCDPs is of potential clinical interest.

Since the algorithm is able to learn the signal characterization from data, we believe that it could also be used for neural spike detection or other signal type classification. However, the focus of the method is on the relation between the peaks in the signal. It requires further investigation to ascertain whether this feature is adequate for other domains.

References

[1] Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91

[2] Musallam S, Corneil B D, Greger B, Scherberger H and Andersen R A 2004 Cognitive control signals for neural prosthetics *Science* **305** 258–62

[3] Lebedev M A and Nicolelis M A 2006 Brain–machine interfaces: past, present an future *Trends Neurosci.* **29** 536–46

[4] Velliste M, Perel S, Spalding M C, Whitford A S and Schwartz A B 2008 Cortical control of a prosthetic arm for self-feeding *Nature* **453** 1098–101

[5] Lotte F, Congedo M, Lécuyer A, Lamarche F and Arnaldi B 2007 A review of classification algorithms for EEG-based brain–computer interfaces *J. Neural Eng.* **4** R1–R13

[6] Bremer F 1941 L'activité électrique 'spontanée' de la moelle épinière *Arch. Physiol. Biochem.* **51** 51–84

[7] ten Cate J 1950 Spontaneous electrical activity of the spinal cord *Electroencephalogr. Clin. Neurophysiol.* **2** 445–51

[8] Kasprzak H and Gasteiger E L 1970 Spinal electrogram of freely moving cat: supraspinal and segmental influences *Brain Res.* **22** 207–20

[9] Rudomin P, Solodkin M and Jiménez I 1987 Synaptic potentials of primary afferent fibers and motoneurons evoked by single intermediate nucleus interneurons in the cat spinal cord *J. Neurophisiol.* **57** 1288–313

[10] Manjárrez E, Jiménez E and Rudomin P 2003 Intersegmental synchronization of spontaneous activity of dorsal horn neurons in the cat spinal cord *Exp. Brain Res.* **148** 401–13

[11] García C A, Chávez D, Jiménez E and Rudomin P 2004 Effects of spinal and peripheral nerve lesions on the intersegmental synchronization of the spontaneous activity of dorsal horn neurons in the cat lumbosacral spinal cord *Neurosci. Lett.* **361** 102–5

[12] Rudomin P 2009 In search of lost presynaptic inhibition *Exp. Brain Res.* **196** 139–51

[13] Chávez D, Rodríguez E, Jiménez I and Rudomin P 2012 Changes in correlation between spontaneous activity of dorsal horn neurones lead to differential recruitment of inhibitory pathways in the cat spinal cord *J. Physiol.* **590** 1563–84

[14] Kaper M, Meinicke P, Grossekathoefer U, Lingner T and Ritter H 2004 BCI competition 2003-data set IIb: support vector machines for the P300 speller paradigm *IEEE Trans. Biomed. Eng.* **51** 1073–6

[15] Pfurtscheller G, Neuper C, Flotzinger D and Pregenzer M 1997 EEG-based discrimination between imagination of right and left hand movement *Electroencephalogr. Clin. Neurophysiol.* **103** 642–51

[16] Chiappa S and Bengio S 2004 HMM and IOHMM modeling of EEG rhythms for asynchronous BCI systems *European Symp. on Artificial Neural Networks* pp 199–204

[17] Penny W D and Roberts S J 1999 EEG-based communication via dynamic neural network models *Int. Joint Conf. on Neural Networks* pp 3586–90

[18] Wang Y, Berg P and Scherg M 1999 Common spatial subspace decomposition applied to analysis of brain responses under multiple task conditions: a simulation study *Clin. Neurophysiol.* **110** 604–14

[19] Qin L, Ding L and He B 2004 Motor imagery classification by means of source analysis for brain–computer interface applications *J. Neural Eng.* **1** 135–41

[20] Vidakovic B 1999 *Statistical Modeling with Wavelets* (New York: Wiley)

[21] Hastie T, Tibshirani R and Friedman J 2008 *The Elements of Statistical Learning: Data Mining, Inference and Predictions* 2nd edn (Berlin: Springer)
[22] Friedman J H, Hastie T and Tibshirani R 2000 Additive logistic regression: a statistical view of boosting *The Ann. Stat.* **28** 337–74
[23] Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Stat.* **29** 1189–232
[24] Jolliffe I T 2002 *Principal Component Analysis* 2nd edn (Berlin: Springer)
[25] Hyvärinen A, Karhunen J and Oja E 2001 *Independent Component Analysis* (New York: Wiley)
[26] Bartholomew D J and Knott M 1999 *Latent Variable Models and Factor Analysis* 2nd edn (London: Hodder Arnold)
[27] Bergman H and DeLong M R 1992 A personal computer-based spike detector and sorter: implementation and evaluation *J. Neurosci. Methods* **41** 187–97
[28] Rudomin P, Hernández E and Lomeli J 2007 Tonic and phasic differential GABAergic inhibition of synaptic actions of joint afferents in the cat *Exp. Brain Res.* **176** 98–118
[29] Katz R 1999 Presynaptic inhibition in humans: a comparison between normal and spastic patients *J. Physiol. (Paris)* **93** 379–85
[30] De Salles A A F, Pedroso A G, Medin P, Agazaryan N, Solberg T, Cabatan-Awang C, Espinosa D M, Ford J and Selch M T 2004 Spinal lesions treated with Novalis shaped beam intensity-modulated radiosurgery and stereotactic radiotherapy *J. Neurosurg.* **101** 435–40